Anticipative Hybrid Extreme Rotation Forest

Borja Ayerdi[1], Manuel Graña[1,2]

[1]Computer Intelligence Group, UPV/EHU, Dept. CCIA, San Sebastian, Spain; [2]ENGINE Centre, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland

ICCS 2016, San Diego, CA, 8th June

Contents

Introduction

Elementary Classifiers

Randomized Data Rotation

Anticipative Hybrid Extreme Rotation Forest

Rationale for AHERF

Experimental design

Experimental Results

Conclusions and future work

Contents

- Introduction
- Elementary Classifiers
- Randomized Data Rotation
- Anticipative Hybrid Extreme Rotation Forest
- Rationale for AHERF
- Experimental design
- Experimental Results
- Conclusions and future work

Overview of the paper

- Adaptive Hybrid Extreme Rotation Forest (AHERF):
  - heterogeneous classifier ensembles
    - profit from classifier specialization
  - the anticipative determination of the the fraction of each classifier architecture included in the ensemble. ,
    - independent pilot classifer architecture cross-validation experiments
    - rank classifier architectures
    - build a probability distribution of classifier architectures
    - type of each individual classifier is decided by sampling

Contents

Introduction

Elementary Classifiers

Randomized Data Rotation

Anticipative Hybrid Extreme Rotation Forest

Rationale for AHERF

Experimental design

Experimental Results

Conclusions and future work

Elementary classifiers

Elementary classifiers implementation in the experiments reported in this paper are extracted from SciKit Python package.

-Decision Trees,

-Extreme Learning Machines

-Support Vector Machines

-k-Nearest Neighbors

-Adaboost

-Gaussian Naive Bayes

The Python implementation of AHERF is available .

Contents

Introduction

Elementary Classifiers

Randomized Data Rotation

Anticipative Hybrid Extreme Rotation Forest

Rationale for AHERF

Experimental design

Experimental Results

Conclusions and future work

Randomized data rotation

To construct the training/testing datasets for a specific classifier D_i in an ensemble, we carry out the following steps:

1. Partition the set of feature variables F into K subsets of variables.
2. For each subset of feature variables, F_k, $k = 1, \ldots, K$
   - 2.1 extract the corresponding data X_k from the training data set
   - 2.2 compute the partial randomized rotation matrix R_k using Principal Component Analysis (PCA) from X_k
3. Compose the global rotation matrix $R = [R_1, \ldots, R_K]$, reordering columns according to the original data,
4. Transform the train and test data applying the same rotation matrix.

Contents

Introduction

Elementary Classifiers

Randomized Data Rotation

Anticipative Hybrid Extreme Rotation Forest

Rationale for AHERF

Experimental design

Experimental Results

Conclusions and future work

Anticipative Hybrid Extreme Rotation Forest

- Let $\mathbf{x} = [x_1, \ldots, x_n]^T$ be a sample described by n feature variables,
- F is the feature variable set and
- X is the data set containing N training samples in a matrix of size $n \times N$.
- Let Y be a vector containing the class labels of the data samples, $Y = [y_1, \ldots, y_N]^T$.
- The number of classes is denoted Ω.
- Denote by $D_1, \ldots, D_L$ the classifiers in the ensemble,

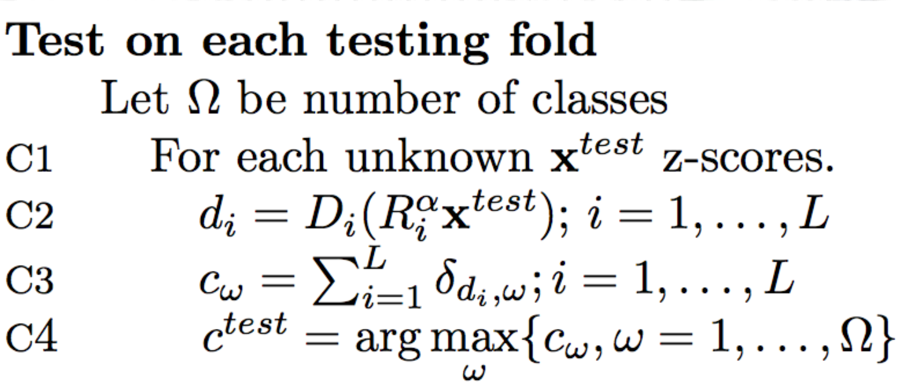
AHERF

```
Begin
Anticipative Model selection
M1   Select 30% of the dataset for model selection
M2   For each classifier type $k = 1, \dots, M$
M3      Perform 5-fold cross-validation, obtain accuracy $A_k$
M4   Rank $A_k$, assigning $r_k$ to the $k$-th classifier
M5   Assign selection probability $p_k = \frac{Fib((C+1)-r_k)}{\sum_{i=1}^{C} Fib(i)}$, $k = 1, \dots, M$

On the 70% unused data, perform 10-fold cv, at each fold:

Ensemble construction on each training fold
2    For each individual classifier $D_i$, $i = 1 \dots L$
3       Computation of rotation matrix $R_i^\alpha$:
4          Partition $F$ into $K$ random subsets: $F_{i,j}; j = 1 \dots K$
5          For each $F_{ij}$, $j = 1 \dots K$
6             - Let $X_{i,j}$ be the subset of $X$ corresponding to features in $F_{i,j}$.
7             - $C_{i,j}$ obtained from PCA on $X_{i,j}$
8          Compose $R_i^\alpha$ using matrices $C_{i,j}$ .
9       Decide the model of $D_i$ sampling $\{p_k; k = 1, \dots, M\}$
10      Train classifier $D_i$ on training set $(R_i^\alpha X, Y)$ or $(X, Y)$
End ensemble construction
```

AHERF



Test on each testing fold

Let Ω be number of classes

C1 For each unknown $\mathbf{x}^{test}$ z-scores.

C2 $d_i = D_i(R_i^{\alpha}\mathbf{x}^{test});\ i = 1, \ldots, L$

C3 $c_{\omega} = \sum_{i=1}^{L} \delta_{d_i,\omega}; i = 1, \ldots, L$

C4 $c^{test} = \arg\max_{\omega}\{c_{\omega}, \omega = 1, \ldots, \Omega\}$

AHERF ranking distribution

- model selection phase uses 30% of the training data
- For each classifier type a 5-fold cross-validation is performed on the selected data.
- r_k is the ranking of the k-th classifier type .
- selection probability according to the expression

$$p_k = \frac{Fib\left((C+1) - r_k\right)}{\sum_{i=1}^{C} Fib\left(i\right)},$$

where $Fib\left(i\right)$ is the i-th value of the Fibonacci series.

AHERF ranking distribution

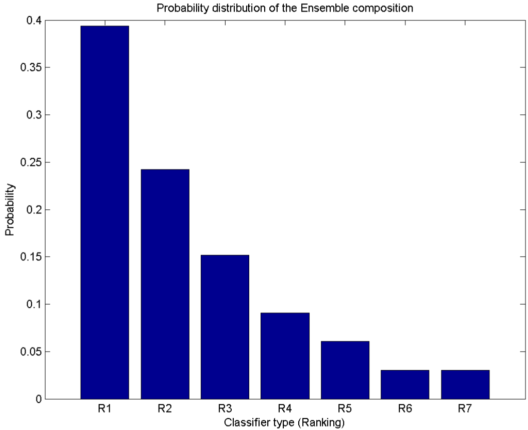


Figure : The architecture selection probability distribution from the ranking of the classifiers.

Contents

Introduction

Elementary Classifiers

Randomized Data Rotation

Anticipative Hybrid Extreme Rotation Forest

Rationale for AHERF

Experimental design

Experimental Results

Conclusions and future work

General Motivation

- Heterogenous ensembles of classifiers are motivated by the well known no-free lunch theorems
    - no single approach is optimal for the solution of all optimization problems,
- it can as well as be applied to machine learning solutions of classification and regression problems.
- Therefore, we would like to predict which kind of classifier architecture is better for the problem domain at hand.
- The idea in AHERF is to build an ensemble where the best fitted classifier types are more frequent.

Some notation

- ground truth classification mapping $\mathcal{C} : \mathcal{X} \to \Omega$,
- that gives the true class $\omega \in \Omega$ corresponding to each input feature vector $\mathbf{x} \in \mathcal{X}$.
- we build classifiers ${}^{t}C$ from $X = \{(\mathbf{x}_i, \omega_i)\}_{i=1}^{N}$,
  - $t \in T$
  - collection of classifier architectures T,
  - its best estimation of the true class $\hat{\omega} =^{t} C\left(\mathbf{x}\right)$.
    - as a maximum *a posteriori* estimation, i.e.

$$\hat{\omega} = \max_{\omega} {}^{t}\hat{P}\left(\omega \left| \mathbf{x} \right.\right),$$

Accuracy

- The accuracy of a classifier can be computed as the expectation of the distance between the a posteriori distribution and the ground truth classification:

$$ {}^{t}A = E_{\mathcal{X}}\left[\left\| \left[{}^{t}\hat{P}\left(\omega \left| \mathbf{x} \right.\right) - \mathcal{C}\left(\omega, \mathbf{x}\right) \right]_{\omega} \right\| \right], $$

  where

  - $E_{\mathcal{X}}\left[.\right]$ denotes the expectation over the input space, i.e. over all possible sampling processes providing the training dataset X, and
  - $\mathcal{C}\left(\omega, \mathbf{x}\right)$ is 1 for the true class, and 0 for the others.

- cross-validation experiments are a minimum variance method to provide estimates of the accuracy.

Acccuracy of the ensemble

- ensemble of classifiers $\{{}^tC_k\}_{k=1}^M$,
- t as many *a posteriori* distribution estimations as classifiers.

$$\left\{\left\{{}^t\hat{P}_k\left(\omega\,|\mathbf{x}\right)\right\}_\omega\right\}_{k=1}^M$$

ensemble decision by majority voting, then the ensemble class estimation is given by

$$\hat{\omega} = \arg\max_\omega \left|\left\{k\,|\omega = \hat{\omega}_k\right\}\right|,$$

where $\hat{\omega}_k = \max_\omega {}^t\hat{P}_k\left(\omega\,|\mathbf{x}\right)$.

- Accuracy of the ensemble can be modeled by

$$A_M \propto E_{\mathcal{X}}\left[\sum_k \left\|\left[{}^t\hat{P}_k\left(\omega\,|\mathbf{x}\right) - \mathcal{C}\left(\omega,\mathbf{x}\right)\right]_\omega\right\|\right]$$

It is immediate that

$$A_M \propto \sum_{k=1}^M \left({}^tA_k\right).$$

Convergence

- Let us assume that there is some accuracy ranking of the classifier types
$$ {}^{t_1}A > {}^{t_2}A > {}^{t_3}A > \dots $$
- an ensemble is characterized by the vector $\mathbf{n} = [n_t \,|\, t \in T^*]$,
  - where T^* denotes the identifiers of the classifiers types ordered by accuracy ranking.
- ensembles can be ordered by lexicographic ordering
  - if $\mathbf{n}' > \mathbf{n}''$ we expect the first ensemble to have accuracy greater than the second.

Convergence

- AHERF estimates the classifier type ranking

  $$\widehat{{}^{t_1}A} > \widehat{{}^{t_2}A} > \widehat{{}^{t_3}A} > ...$$

  using this information to drive the selection of the classifier type of each individual ensemble constituent.
- In order to have ensembles whose characteristic vector **n** is of the form

  $$n_{t_1} >> n_{t_2} >> n_{t_3} > ...$$

  we sample an integer random variable whose distribution of probability is an approximation of the exponential distribution built using the Fibbonacci series on the ranking.

Contents

Introduction

Elementary Classifiers

Randomized Data Rotation

Anticipative Hybrid Extreme Rotation Forest

Rationale for AHERF

Experimental design

Experimental Results

Conclusions and future work

Experimental design

- Validation
  - the average of 50 repetitions of a 10-fold cross-validation approach,
- all feature extraction and classification parameters are estimated from the training datasets and applied to the testing datasets as such.
- data normalization by the independent computation of the z-score of each input variable
  - the μ and σ are estimated on the training data and used as such on the testing data,

Experimental design

Model parameter selection

- L: The number of individual classifiers,is set to $L = 35$ for all experiments.
- Classifier intrinsic parameters:
- DT depth is set to 10 i
- The number of hidden nodes in the ELM is set to $\min\left\{\frac{N}{3}, 1000\right\}$.
- The SFLN architecture trained by ELM has a single output unit encoding the output of the classifier as an integer value, both for two-class and many-classes datasets.
- K: The number of partitions of the set of features has been set to $K = \left\lfloor \frac{n}{4} \right\rfloor$.

Materials

We have performed the computational experiments over 16 datasets used for the comparison and validation are in the public domain, they have been extracted from the UCI machine learning repository [1], including multi-class instances as well as two class problems.

[1] http://archive.ics.uci.edu/ml/

Contents

Introduction

Elementary Classifiers

Randomized Data Rotation

Anticipative Hybrid Extreme Rotation Forest

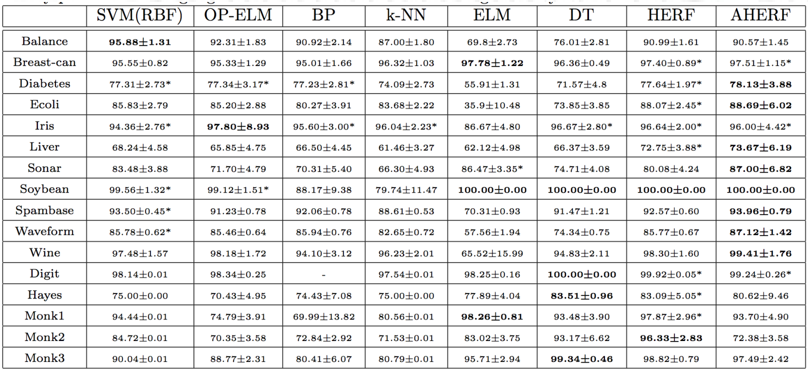
Rationale for AHERF

Experimental design

Experimental Results

Conclusions and future work

Experimental results



	SVM(RBF)	OP-ELM	BP	k-NN	ELM	DT	HERF	AHERF
Balance	**95.88±1.31**	92.31±1.83	90.92±2.14	87.00±1.80	69.8±2.73	76.01±2.81	90.99±1.61	90.57±1.45
Breast-can	95.55±0.82	95.33±1.29	95.01±1.66	96.32±1.03	**97.78±1.22**	96.36±0.49	97.40±0.89*	97.51±1.15*
Diabetes	77.31±2.73*	77.34±3.17*	77.23±2.81*	74.09±2.73	55.91±1.31	71.57±4.8	77.64±1.97*	**78.13±3.88**
Ecoli	85.83±2.79	85.20±2.88	80.27±3.91	83.68±2.22	35.9±10.48	73.85±3.85	88.07±2.45*	**88.69±6.02**
Iris	94.36±2.76*	**97.80±8.93**	95.60±3.00*	96.04±2.23*	86.67±4.80	96.67±2.80*	96.64±2.00*	96.00±4.42*
Liver	68.24±4.58	65.85±4.75	66.50±4.45	61.46±3.27	62.12±4.98	66.37±3.59	72.75±3.88*	**73.67±6.19**
Sonar	83.48±3.88	71.70±4.79	70.31±5.40	66.30±4.93	86.47±3.35*	74.71±4.08	80.08±4.24	**87.00±6.82**
Soybean	99.56±1.32*	99.12±1.51*	88.17±9.38	79.74±11.47	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**
Spambase	93.50±0.45*	91.23±0.78	92.06±0.78	88.61±0.53	70.31±0.93	91.47±1.21	92.57±0.60	**93.96±0.79**
Waveform	85.78±0.62*	85.46±0.64	85.94±0.76	82.65±0.72	57.56±1.94	74.34±0.75	85.77±0.67	**87.12±1.42**
Wine	97.48±1.57	98.18±1.72	94.10±3.12	96.23±2.01	65.52±15.99	94.83±2.11	98.30±1.60	**99.41±1.76**
Digit	98.14±0.01	98.34±0.25	-	97.54±0.01	98.25±0.16	**100.00±0.00**	99.92±0.05*	99.24±0.26*
Hayes	75.00±0.00	70.43±4.95	74.43±7.08	75.00±0.00	77.89±4.04	**83.51±0.96**	83.09±5.05*	80.62±9.46
Monk1	94.44±0.01	74.79±3.91	69.99±13.82	80.56±0.01	**98.26±0.81**	93.48±3.90	97.87±2.96*	93.70±4.90
Monk2	84.72±0.01	70.35±3.58	72.84±2.92	71.53±0.01	83.02±3.75	93.17±6.62	**96.33±2.83**	72.38±3.58
Monk3	90.04±0.01	88.77±2.31	80.41±6.07	80.79±0.01	95.71±2.94	**99.34±0.46**	98.82±0.79	97.49±2.42

Results discussion

- It can be appreciated that AHERF gives the best results in most cases
  - (Ecoli: 88.69%; Liver: 73.67%; Sonar: 87%; Spambase: 93.96%, etc)
- and it is close to the best result in the others.
- Differences are not statistically significant (t-test p>0.01) due to high variance of the results

algorithm working

- we show
  - an instance of the ranking of the classifier types for each database, and
  - the number of individual classifiers of each type generated by selection according to those rankings.
- there is no guarantee that the better ranking will lead to a greater number of individual classifiers in the ensemble, due to random nature of the generation process,
- AHERF is better suited for big datasets.

Results

Table : Ranking (1-best, 7-worst) of elementary classifier types per each benchmark database.

	DT	ELM	k-NN	SVM (RBF)	RF	AdaBoost	Gaussian NB
Balance	6	5	2	4	7	1	3
Breast-can	5	3	4	2	6	7	1
Diabetes	2	6	5	1	4	7	3
Ecoli	6	2	5	4	3	7	1
Iris	6	7	5	4	3	2	1
Liver	6	1	7	5	4	3	2
Sonar	6	7	3	2	5	4	1
Soybean	6	7	5	4	3	2	1
Spambase	5	4	6	2	3	1	7
Waveform	6	7	3	1	2	4	5
Wine	6	5	1	3	4	7	2
Digit	2	4	6	5	3	7	1
Hayes	2	6	5	4	1	7	3

Results

Table : Number of classifiers on an instance of final ensemble composition

	DT	ELM	k-NN	SVM (RBF)	RF	AdaBoost	Gaussian NB
Balance	6	1	4	2	3	15	4
Breast-can	1	3	4	7	1	3	16
Diabetes	7	1	2	19	1	1	4
Ecoli	2	6	3	2	9	0	13
Iris	1	0	5	3	4	10	12
Liver	3	10	0	1	3	10	8
Sonar	0	2	2	9	6	4	12
Soybean	0	0	3	5	4	9	14
Spambase	4	2	1	9	7	10	2
Waveform	0	2	4	18	10	0	1
Wine	0	2	16	3	3	0	11
Digit	10	0	1	4	5	3	12
Hayes	5	2	3	5	11	3	6

Contents

Introduction

Elementary Classifiers

Randomized Data Rotation

Anticipative Hybrid Extreme Rotation Forest

Rationale for AHERF

Experimental design

Experimental Results

Conclusions and future work

Conclusions

- The proposal of the AHERF hybrid ensemble classifier is an improvement of HERF algorithm, including the anticipative selection of the classifier type according to the prediction of the classifier types accuracy in each database.
- The results obtained on a collection of benchmark databases are encouraging.
- Further works
  - to apply AHERF in other areas like medical image processing (fMRI, CTA, etc) and remote sensing image processing problems, and
  - to improve the combination of the outputs of the ensemble.